# Combining Spatial Statistical and Ensemble Information in Probabilistic Weather Forecasts

Veronica J. Berrocal, Adrian E. Raftery, and Tilmann Gneiting

Technical Report no. 495
Department of Statistics
University of Washington

February 22, 2006

# Report Documentation Page

*Form Approved*
*OMB No. 0704-0188*

1. REPORT DATE
**22 FEB 2006**

2. REPORT TYPE

3. DATES COVERED
**00-02-2006 to 00-02-2006**

4. TITLE AND SUBTITLE
**Combining Spatial Statistical and Ensemble Information in Probabilistic Weather Forecasts**

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
**University of Washington,Department of Statistics,Box 354322,Seattle,WA,98195-4322**

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
**Approved for public release; distribution unlimited**

13. SUPPLEMENTARY NOTES
**The original document contains color images.**

14. ABSTRACT

15. SUBJECT TERMS

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | **31** | |
| **unclassified** | **unclassified** | **unclassified** | | | |

## Abstract

Forecast ensembles typically show a spread-skill relationship, but they are also often underdispersive, and therefore uncalibrated. Bayesian model averaging (BMA) is a statistical postprocessing method for forecast ensembles that generates calibrated probabilistic forecast products for weather quantities at individual sites. This paper introduces the Spatial BMA technique, which combines BMA and the geostatistical output perturbation (GOP) method, and extends BMA to generate calibrated probabilistic forecasts of whole weather fields simultaneously, rather than just weather events at individual locations. At any site individually, Spatial BMA reduces to the original BMA technique.

The Spatial BMA method provides statistical ensembles of weather field forecasts that take the spatial structure of observed fields into account and honor the flow-dependent information contained in the dynamical ensemble. The members of the Spatial BMA ensemble are obtained by dressing the weather field forecasts from the dynamical ensemble with simulated spatially correlated error fields, in proportions that correspond to the BMA weights for the member models in the dynamical ensemble. Statistical ensembles of any size can be generated at minimal computational costs.

The Spatial BMA technique was applied to 48-h forecasts of surface temperature over the North American Pacific Northwest in 2004, using the University of Washington mesoscale ensemble. The Spatial BMA ensemble generally outperformed the BMA and GOP ensembles and showed much better verification results than the raw ensemble, both at individual sites, for weather field forecasts, and for forecasts of composite quantities, such as average temperature in National Weather Service forecast zones and minimum temperature along the Interstate 90 Mountains to Sound Greenway.

# Contents

# List of Tables

# List of Figures

# 1 Introduction

Ensemble prediction systems have been developed to generate probabilistic forecasts of weather quantities that address the two major sources of forecast uncertainty in numerical weather prediction: uncertainty in initial conditions, and uncertainty in model formulations. Originally suggested by Epstein (1969) and Leith (1974), ensemble forecasts have been operationally implemented on the synoptic scale (Toth and Kalnay 1993; Houtekamer et al. 1996; Molteni et al. 1996) and are under development on the mesoscale (Stensrud et al. 1999; Wandishin et al. 2001; Grimit and Mass 2002; Eckel and Mass 2005). In a wide range of applications, probabilistic forecasts based on ensembles provide higher economic and societal value than a single deterministic forecast (Richardson 2000; Palmer 2002; Gneiting and Raftery 2005).

While showing significant spread–error correlations, ensemble forecasts are often biased and underdispersive (Buizza 1997; Hamill and Colucci 1997; Grimit and Mass 2002; Scherrer et al. 2004; Eckel and Mass 2005). Hence, to realize the full potential of an ensemble forecast it is necessary to apply some form of statistical postprocessing, with the goal of generating probabilistic forecasts that are calibrated and yet sharp. In the spirit of the pioneering work of Glahn and Lowry (1972), who introduced regression-type model output statistics approaches to a meteorological audience, various statistically-based ensemble postprocessing techniques have been proposed. In this paper, we introduce a postprocessing technique that combines two of these methods, Bayesian model averaging (Raftery et al. 2005) and the geostatistical output perturbation technique (Gel et al. 2004a), to generate calibrated probabilistic forecasts of whole weather fields simultaneously, rather than just weather quantities at individual locations.

Bayesian model averaging (BMA) is a statistical technique originally developed for social and health science applications in situations with several competing statistical models (Hoeting et al. 1999). Raftery et al. (2005) proposed the use of BMA to calibrate forecast ensembles and generate predictive probability density functions (PDFs) for future weather quantities. The BMA predictive PDF is a weighted average of predictive PDFs associated with each individual ensemble member, with weights that reflect the member's relative skill. However, each location in the forecast domain is considered individually, and spatial correlations among forecast errors are ignored.

The geostatistical output perturbation (GOP) method dresses a single deterministic weather field forecast with simulated error fields, to obtain statistical ensembles of weather fields that take spatial correlations into account (Gel et al. 2004a). This resembles the perturbation approach in Houtekamer and Mitchell (1998, 2001), but in the GOP technique spatially correlated perturbations are applied to the outputs of numerical weather prediction models, rather than the inputs.

In essence, the BMA technique honors ensemble information but ignores spatial correlation. The GOP method takes spatial dependencies into account, but applies to a single deterministic forecast, rather than an ensemble of weather field forecasts, and fails to honor the flow-dependent spread that derives from the nonlinear evolution of the atmosphere and is characteristic for dynamical ensembles.

Spatial BMA addresses these shortcomings by combining the two techniques. As in the original BMA technique, the Spatial BMA predictive PDF is a weighted average of forecast PDFs centered at bias-corrected versions of the ensemble member models, with weights that relate to each member's performance. However, in Spatial BMA the forecast PDFs are multivariate densities with covariance structures designed to honor the spatial structure of weather observations. The Spatial BMA technique can be used to generate statistical ensembles of whole weather fields simultaneously, of any size, and at minimal computational costs. At any location individually, Spatial BMA reduces to the original BMA technique.

The paper is organized as follows. In Section 2, we review the BMA and GOP methods and describe the Spatial BMA technique in detail. In Section 3 we give an example of Spatial BMA forecasts of surface temperature over the North American Pacific Northwest, using the University of Washington mesoscale ensemble (Grimit and Mass 2002; Eckel and Mass 2005). Section 4 presents verification results for Spatial BMA forecasts in calendar year 2004, focusing on spatial and composite quantities. The paper ends with a discussion in Section 5, in which we compare the Spatial BMA technique to the dressing approaches of Roulston and Smith (2003) and Wang and Bishop (2005).

# 2 Methods

We now describe the BMA, GOP and Spatial BMA techniques, and we explain our approach to parameter estimation.

## 2.1 Bayesian model averaging (BMA)

We consider an ensemble of $K$ weather field forecasts. In our examples, this is the eight-member University of Washington mesoscale ensemble (UWME; Eckel and Mass 2005), but BMA applies to all forecast ensembles with physically distinguishable member models, such as poor person's or multi-model ensembles. With small modifications, BMA also applies to ensembles with exchangeable members, including bred and singular-vector ensembles

(Raftery et al. 2005).

We write $y(s)$ for the weather quantity of interest at location $s$, and $f_1(s), \ldots, f_K(s)$ for the respective ensemble member forecasts. With each ensemble member, we associate a conditional PDF, $g_k(y(s) \,|\, f_k^0(s))$, which we interpret as the conditional PDF of $y(s)$ given that member $k$ is the best among the ensemble member forecasts, as indicated by the superscript. The BMA predictive PDF for the weather quantity at site $s$ then is

$$p\,(y(s) \,|\, f_1(s), \ldots, f_k(s)) = \sum_{k=1}^{K} w_k \, g_k(y(s) \,|\, f_k^0(s)), \tag{1}$$

where $w_k$ is the probability of ensemble member $k$ being best. In the implementation of Raftery et al. (2005), which applies to forecasts of surface temperature and sea-level pressure, the conditional PDFs are univariate normal densities centered at a linearly bias-corrected forecast. Hence, $g_k(y(s) \,|\, f_k^0(s))$ is a univariate normal PDF with mean $a_k + b_k f_k(s)$ and standard deviation $\sigma_0$, assumed to be constant across ensemble members. We denote this situation by

$$y(s) \,|\, f_k^0(s) \sim \mathcal{N}(a_k + b_k f_k(s), \sigma_0^2). \tag{2}$$

The BMA weights in (1) and the bias and variance parameters in (2) are estimated from training data. The BMA weights reflect the relative performance of the ensemble member models during the training period; since they are probabilities, they are nonnegative and their sum is equal to 1.

The BMA method as specified by (1) is implemented in the EnsembleBMA package for the R language (Ihaka and Gentleman 1996), which is available for download at http://cran.r-project.org.

## 2.2 The geostatistical output perturbation (GOP) technique

The geostatistical output perturbation (GOP) technique dresses a single deterministic weather field forecast with Gaussian error fields that are generated using geostatistical methods (Gel et al. 2004a). Here, we take the deterministic weather field forecast to be a member of the dynamical ensemble.

Specifically, let $S$ denote a possibly large but finite set of distinct model gridpoints or scattered observation sites. If our intention is to produce stamp maps of weather field forecasts, this set is the model grid. For verification purposes, it is a collection of observation locations, and the forecasts are bilinearly interpolated from the model grid to the observation sites. We write

$$\mathbf{Y} = \{y(s) : s \in S\}$$

for the weather field at the sites of interest, and $\mathbf{F}_k = \{f_k(s) : s \in S\}$ for the respective deterministic weather field forecast. The GOP technique employs a statistical model which assumes that

$$\mathbf{Y}|\mathbf{F}_k \sim \mathcal{MVN}(a_k + b_k\mathbf{F}_k, \Sigma_k), \tag{3}$$

where the right-hand side denotes a multivariate normal PDF centered at the bias-corrected member forecast, $a_k+b_k\mathbf{F}_k$, and with covariance matrix $\mathbf{\Sigma}_k$, with entries specified in (4) below. Superficially, one might think of (3) as a spatial version of (2), but the relationships differ fundamentally: in (3), we consider $\mathbf{F}_k$ as a single deterministic forecast without reference to any of the other ensemble members; in (2), we consider $f_k(s)$ conditionally on this member being the best among the ensemble member forecasts. This latter assumption of forecast $k$ being best generally implies a deflated variance in (2), when compared to (3), as will be seen below. For surface temperature and sea level pressure, the use of a multivariate normal PDF seems reasonable as an approximation, but this may not be true for other weather variables, such as precipitation or wind speed.

The covariance matrix in (3) describes the spatial structure of the forecast error field and needs to be estimated from training data. Gel et al. (2004a) used a parametric, stationary and isotropic geostatistical model, which assumes that the $(i, j)$th element of the covariance matrix $\mathbf{\Sigma}_k$ is

$$\rho_k^2\,\delta_{ij} + \tau_k^2\,\exp\!\left(-\frac{\|s_i - s_j\|}{r_k}\right), \tag{4}$$

where $\|s_i - s_j\|$ denotes the Euclidean distance between the respective locations, $s_i$ and $s_j$, and $\delta_{ij}$ equals 1 if $s_i = s_j$ and is 0 otherwise. In geostatistical terminology, $\rho_k^2$ is called the nugget effect, $\rho_k^2 + \tau_k^2$ is known as the sill, and $r_k$ is called the range and indicates the rate at which the spatial correlations of the forecast errors decay (Cressie 1993; Chilès and Delfiner 1999). Covariance structures that are more complex can be accommodated, and we discuss some of the options in Section 5.

Note that (3) and (4) give a fully specified, multivariate normal predictive PDF for the weather field $\mathbf{Y}$. To generate statistical ensembles from this PDF, we express (3) and (4) in the form of the stochastic representation

$$\mathbf{Y}|\mathbf{F}_k \sim a_k + b_k\mathbf{F}_k + \mathbf{E}_{1k} + \mathbf{E}_{2k}, \tag{5}$$

where $\mathbf{F}_k$ is the deterministic weather field forecast, $a_k$ and $b_k$ are scalar bias parameters, and $\mathbf{E}_{1k} = \{\epsilon_{1k}(s) : s \in S\}$ and $\mathbf{E}_{2k} = \{\epsilon_{2k}(s) : s \in S\}$ are independent random vectors with mean zero, satisfying

$$\text{cov}\,(\epsilon_{1k}(s_i), \epsilon_{1k}(s_j)) = \tau_k^2\,\exp\!\left(-\frac{\|s_i - s_j\|}{r_k}\right),$$

and $\text{cov}(\epsilon_{2k}(s_i), \epsilon_{2k}(s_j)) = \rho_k^2 \, \delta_{ij}$, respectively. In this representation, $\mathbf{E}_{1k}$ is a spatially correlated error field that varies continuously with distance, and we refer to it as the continuous component of the forecast error field. In contrast, $\mathbf{E}_{2k}$ is a noise vector that stands for instrument and representativeness errors, and we refer to it as the discontinuous component of the error field. Statistical GOP ensembles of any size can be obtained by simulating $\mathbf{E}_{1k}$ and $\mathbf{E}_{2k}$ from their respective multivariate PDFs, and adding the simulated errors to the bias-corrected forecast, as directed by (5). For the simulations, we use the circulant embedding technique (Wood and Chan 1994; Gneiting et al. 2006) as implemented in the RandomFields package for the R language (Schlather 2001). The GOP method is itself implemented in the ProbForecastGOP package for the R language. All R packages are available for download at http://cran.r-project.org.

## 2.3 Spatial BMA

We now show how to combine the BMA and GOP methods into the Spatial BMA technique. Again, we consider a weather field $\mathbf{Y} = \{Y(s) : s \in S\}$ at a possibly large but finite collection $S$ of locations, but now conditionally on an ensemble,

$$\mathbf{F}_1 = \{f_1(s) : s \in S\}, \; \ldots \;, \mathbf{F}_K = \{f_K(s) : s \in S\}$$

of $K$ weather field forecasts simultaneously, rather than just a single deterministic weather field forecast. The Spatial BMA predictive PDF for the weather field is

$$p\,(\mathbf{Y}|\mathbf{F}_1, \ldots, \mathbf{F}_K) = \sum_{k=1}^{K} w_k \, g_k(\mathbf{Y}|\mathbf{F}_k^0), \tag{6}$$

where $w_k$ is the respective BMA weight and is the probability that member $k$ is the best among the ensemble member forecasts, and $g_k(\mathbf{Y}|\mathbf{F}_k^0)$ is the conditional PDF of $\mathbf{Y}$ given that member $k$ is best, as indicated by the superscript. In our implementation, the conditional PDFs are multivariate normal densities centered at the bias-corrected ensemble member forecast, $a_k + b_k\mathbf{F}_k$, and having a spatially structured covariance matrix, $\mathbf{\Sigma}_k^0$. By analogy to (2), we denote this situation by

$$\mathbf{Y}|\mathbf{F}_k^0 \sim \mathcal{MVN}\left(a_k + b_k\mathbf{F}_k, \mathbf{\Sigma}_k^0\right). \tag{7}$$

In (7),

$$\mathbf{\Sigma}_k^0 = \frac{\sigma_0^2}{\rho_k^2 + \tau_k^2} \, \Sigma_k, \tag{8}$$

where $\sigma_0^2$ is the BMA variance in (2), $\mathbf{\Sigma}_k$ the spatially structured GOP covariance matrix with entries specified in (4), and $\rho_k^2$ and $\tau_k^2$ the respective GOP covariance parameters. The quantity

$$\alpha_k = \frac{\sigma_0^2}{\rho_k^2 + \tau_k^2}$$

is the ratio of the BMA variance to the GOP variance for the forecast errors, and we refer to it as the deflation factor associated with member model $k$, where $k = 1, \ldots, K$. Spatial BMA generalizes both the standard BMA method and the GOP technique: It reduces to the former when the set $S$ consists of a single location only, and it reduces to the latter for an ensemble of size $K = 1$, that is, a deterministic weather field forecast.

Similarly to GOP, the Spatial BMA equations (6) through (8) give a fully specified, multivariate predictive PDF for the weather field. However, it is more practical to generate a statistical ensemble of weather field forecasts, by sampling from the Spatial BMA predictive PDF. Conditionally on ensemble member $k$ being best, we can write (7) as

$$\mathbf{Y} \,|\, \mathbf{F}_k^0 \sim a_k + b_k \mathbf{F}_k + \mathbf{E}_{1k}^0 + \mathbf{E}_{2k}^0, \tag{9}$$

where $\mathbf{E}_{1k}^0$ and $\mathbf{E}_{2k}^0$ denote the continuous and the discontinuous parts of the conditional forecast error field, respectively, with multivariate normal PDFs equal to those described in Section 2.2 for the unconditional counterparts, $\mathbf{E}_{1k}$ and $\mathbf{E}_{2k}$, except that the covariance matrix is rescaled by the deflation factor, $\alpha_k$.

The following algorithm generates a member of the Spatial BMA ensemble:

1. Sample a number $k \in \{1, \ldots, K\}$, with probabilities given by the BMA weights, $w_1, \ldots, w_K$. This specifies the member of the dynamical ensemble to be dressed.

2. Simulate realizations of the continuous and discontinuous parts, $\mathbf{E}_{1k}^0$ and $\mathbf{E}_{2k}^0$, of the conditional forecast error field from the respective conditional PDFs.

3. Use the right-hand side of (9) to dress the bias-corrected weather field forecast, $a_k + b_k \mathbf{F}_k$, with the simulated conditional forecast error fields, $\mathbf{E}_{1k}^0$ and $\mathbf{E}_{2k}^0$.

Proceeding in this manner, we obtain Spatial BMA ensembles of weather field forecasts, of any desired ensemble size, and at minimal computational cost.

## 2.4 Parameter estimation

The estimation of a Spatial BMA model for an underlying dynamical ensemble requires the fitting of a BMA model as well as GOP models for the individual ensemble members. This is done using prior observations and ensemble forecasts for the same prediction horizon and forecast cycle, with forecasts that are bilinearly interpolated from the model grid to the observation sites. We use a sliding training period consisting of the recent past. In choosing the width of this training period, there is a trade off, in that short training periods allow to better adapt to changes in the ensemble and in its component members, as well as to seasonal changes, while longer training periods tend to decrease estimation variability. Raftery et al. (2005) showed that for 48-hour BMA forecasts of surface temperature in the North American Pacific Northwest there are substantial gains in increasing the length of the training period to 25 days, but there is little gain beyond. In the examples below, we adopt this choice of a sliding 25-day training period. Other weather variables, domains and forecast lead times may require different choices.

To fit the BMA model (1) and (2), we follow Raftery et al. (2005) in estimating the bias parameters, $a_k$ and $b_k$, by linear least squares regression of the observations on the respective ensemble member forecast. The BMA weights, $w_k$, and the BMA variance, $\sigma_0^2$, are estimated using the maximum likelihood technique in the form of the EM algorithm (Dempster et al. 1977) with a subsequent minimum CRPS step.

It remains to fit the GOP models for the weather field forecasts using member model $k$, where $k = 1, \ldots, K$. In estimating the spatial covariance parameters, $\rho_k^2$, $\tau_k^2$ and $r_k$ in (4), it is convenient to note that the GOP error field, $\epsilon_k(s) = \epsilon_{1k}(s) + \epsilon_{2k}(s)$, satisfies

$$\frac{1}{2}\,\mathrm{E}\left(\epsilon_k(s_i) - \epsilon_k(s_j)\right)^2 = \rho_k^2 + \tau_k^2 \left(1 - \exp\left(-\frac{\|s_i - s_j\|}{r_k}\right)\right),$$

where E denotes the expectation operator. In geostatistical language, the error field has variogram

$$\gamma_k(d) = \rho_k^2 + \tau_k^2 \left(1 - e^{-d/r_k}\right),$$

where $d = \|s_i - s_j\|$ denotes the Euclidean distance between two distinct observation sites, and $\gamma_k(d)$ is one-half the expected squared difference between errors at stations that are distance $d$ apart.

We now compute the sample version of the variogram, $\widehat{\gamma}_k(d)$, using data from the sliding training period, as follows:

1. For each day in the training period, find the empirical error field, by subtracting the bias-corrected forecast field from the respective field of verifying weather observations.

2. For each day in the training period, and for all pairs of observation locations on that day, find the distance between the sites, and compute one-half the squared difference between the forecast errors.

3. Group the distances into bins $B_l$ with midpoints $d_l$.

4. Compute the empirical variogram value $\widehat{\gamma}_k(d_l)$ at distance $d_l$, by averaging the respective one-half squared differences over the distance bin $B_l$.

With this, we apply the weighted least squares technique to estimate the GOP parameters. Specifically, if $n_l$ denotes the total number of pairs of observation sites whose distance falls into bin $B_l$, the weighted least squares estimates of the covariance parameters $\rho_k^2$, $\tau_k^2$ and $r_k$ are the values that minimize

$$S\Big(\rho_k^2, \tau_k^2, r_k\Big) = \sum_l n_l \left( \frac{\widehat{\gamma}_k(x_l) - (\rho_k^2 + \tau_k^2 \, (1 - e^{-d_l/r_k}))}{\rho_k^2 + \tau_k^2 \, (1 - e^{-d_l/r_k})} \right)^2 .$$

To solve this optimization problem, we use the quasi-Newton and conjugate-gradient techniques described by Byrd et al. (1995) and implemented in the R language (Ihaka and Gentleman 1996).

Following the estimation of the spatial covariance parameters for ensemble members $k = 1, \ldots, K$, we combine the GOP and BMA models into the Spatial BMA model, using (6) through (8). We do the estimation using the previously mentioned R packages, ensembleBMA and ProbForecastGOP.

# 3 Example

We now give an example of 48-h Spatial BMA forecasts of surface temperature over the North American Pacific Northwest, using the University of Washington mesoscale ensemble (UWME; Grimit and Mass 2002; Eckel and Mass 2005). In the 2004 version used here, the UWME is an eight-member multianalysis ensemble. The members use the fifth-generation Pennsylvania State University–National Center for Atmospheric Research (PSU-NCAR) Mesoscale Model (MM5) driven by initial and lateral boundary conditions supplied by eight distinct global models. Specifically; the **avn** member uses initial and lateral boundary conditions from the Global Forecast System run by the US National Centers for Environmental Prediction (NCEP); the **cmcg** member is based on the Global Environmental Multi-Scale model run by the Canadian Meteorological Centre; the **eta** member uses the limited-area mesoscale model run by NCEP; the **gasp** member is based on the Global AnalysiS and Prediction model run by the Australian Bureau of Meteorology; the **jma** member is based on the Global Spectral Model run by the Japan Meteorological Agency; the **ngps** member uses the Navy Operational Global Atmospheric Prediction System run by the Fleet Numerical Meteorology and Oceanography Center; the **tcwb** member is based on the Global Forecast System run by the Taiwan Central Weather Bureau; and the **ukmo** member derives from the Unified Model run by the UK Met Office. Eckel and Mass (2005) give a detailed description of UWME.

Our example is for 48-h forecasts of the surface (2-m) temperature field over the North American Pacific Northwest, initialized at 0000 UTC on 14 February 2004. To deal with

Table 1: Estimates of BMA parameters for 48-h forecasts of surface temperature verifying at 0000 UTC on 16 February 2004, using UWME. The unit for the standard deviation $\sigma_0$ is degrees Celsius.

| | Land | | | | Ocean | | | |
|---|---|---|---|---|---|---|---|---|
| Member | $w_k$ | $a_k$ | $b_k$ | $\sigma_0^2$ | $w_k$ | $a_k$ | $b_k$ | $\sigma_0^2$ |
| avn | 0.11 | 0.93 | 0.90 | 7.78 | 0.03 | 1.13 | 0.87 | 5.20 |
| cmcg | 0.12 | 0.97 | 0.88 | 7.78 | 0.49 | 1.21 | 0.86 | 5.20 |
| eta | 0.19 | 1.05 | 0.91 | 7.78 | 0.08 | 1.23 | 0.86 | 5.20 |
| gasp | 0.00 | 0.88 | 0.87 | 7.78 | 0.00 | 1.05 | 0.87 | 5.20 |
| jma | 0.15 | 0.98 | 0.92 | 7.78 | 0.05 | 1.17 | 0.89 | 5.20 |
| ngps | 0.27 | 1.04 | 0.90 | 7.78 | 0.15 | 1.18 | 0.87 | 5.20 |
| tcwb | 0.00 | 0.85 | 0.83 | 7.78 | 0.00 | 1.08 | 0.83 | 5.20 |
| ukmo | 0.16 | 0.97 | 0.88 | 7.78 | 0.20 | 1.14 | 0.86 | 5.20 |

nonstationarities in the forecast error fields, we divided the 12-km UWME forecast grid into two subdomains, land and ocean, and estimated separate Spatial BMA models for the two domains, using the aforementioned 25-day sliding training period.

Table 1 shows estimates of the BMA variance, $\sigma_0^2$, the BMA weights, $w_k$, and the additive and multiplicative bias, $a_k$ and $b_k$, respectively, for the eight UWME members. The BMA weights differed substantially between land and ocean. The member with the highest BMA weight on land was the ngps model, and the cmcg model had the highest weight over the ocean. The gasp and tcwb models performed poorly relative to the other members during the training period and received negligible weights in both domains. The two domains also differed in terms of the BMA variance, which was smaller over the Pacific Ocean, likely because of a decrease in the representativeness error.

Table 2 shows estimates of the GOP covariance parameters, $\rho_k^2, \tau_k^2$ and $r_k$, for the forecast error fields, along with estimates of the deflation factor, $\alpha_k$. The estimates of the nugget effect, $\rho_k^2$, which subsumes instrument and representativeness errors, were much larger on land than over the Pacific Ocean. The estimates of $\tau_k^2$ were generally somewhat larger on land than over ocean. The range, $r_k$, corresponds to the correlation length of the continuous component of the error field, with spatial correlations decaying to about 0.05 at distance $3r_k$. The estimates of the range were larger over the ocean than on land, suggesting stronger correlations over water.

The deflation factor, $\alpha_k$, reflects the skill of each ensemble member, with the more accurate members receiving the higher estimates. Indeed, if a member model generally performs well, then the conditional error variance, given that it is best among the ensemble member

Table 2: Estimates of Spatial BMA covariance parameters and deflation factors for 48-h forecasts of surface temperature verifying 0000 UTC on 16 February 2004, using UWME. The unit for the standard deviations $\rho_k$ and $\tau_k$ is degrees Celsius, and the unit for $r_k$ is km.

| | Land | | | | Ocean | | | |
|---|---|---|---|---|---|---|---|---|
| Member | $\rho_k^2$ | $\tau_k^2$ | $r_k$ | $\alpha_k$ | $\rho_k^2$ | $\tau_k^2$ | $r_k$ | $\alpha_k$ |
| avn | 2.26 | 6.30 | 129 | 0.91 | 1.08 | 5.87 | 258 | 0.75 |
| cmcg | 2.32 | 6.06 | 134 | 0.93 | 1.07 | 5.10 | 246 | 0.84 |
| eta | 2.24 | 6.08 | 124 | 0.94 | 1.06 | 5.58 | 245 | 0.78 |
| gasp | 2.31 | 7.25 | 163 | 0.81 | 1.02 | 6.11 | 265 | 0.73 |
| jma | 2.29 | 6.24 | 134 | 0.91 | 1.12 | 5.96 | 277 | 0.73 |
| ngps | 2.20 | 5.37 | 105 | 1.03 | 1.05 | 5.16 | 245 | 0.84 |
| tcwb | 2.35 | 6.67 | 149 | 0.86 | 1.03 | 6.98 | 312 | 0.65 |
| ukmo | 2.29 | 6.39 | 141 | 0.90 | 0.98 | 5.29 | 211 | 0.83 |

forecasts, will not be very different from the unconditional error variance, and the deflation factor will be close to 1. Still, caution is needed in interpreting estimates of deflation factors. For instance, the estimated deflation factors in Table 2 were generally higher on land than they were over the ocean, and the land deflation factor for the ngps model was larger than 1, counter to intuition. These patterns can be explained by Figure 1, which illustrates the estimation of the GOP covariance parameters for the ngps member on land and the cmcg member over the ocean. Each panel shows both the empirical variogram of the forecast error field, composited over the training period, and the fitted exponential variogram. The intercept of the fitted exponential variogram equals the estimate of the nugget effect, $\rho_k^2$, and corresponds to the variance of instrument and representativeness errors. The horizontal asymptote is at the estimated sill, $\rho_k^2 + \tau_k^2$, and equals the estimated marginal variance of the GOP error field. The weighted least squares technique seems to underestimate the sill for the ngps member on land, resulting in a deflation factor that exceeds 1.

The exponential variograms fit quite well over the first 400 km, and the fit deteriorates thereafter. This is quite typical of geostatistical applications, and is not a matter of concern. Generally, when fitting a parametric variogram model, attention is focused on the smaller distances, which are particularly relevant in characterizing the spatial statistical properties of the error fields.

Figures 2 and 3 illustrate the generation of a member of the Spatial BMA ensemble on land and over the ocean, respectively. In each figure, panel (a) shows the bias-corrected member of the dynamical ensemble that is to be dressed. On land, this is the ngps member, and over the ocean it is the cmcg member. Panels (b) and (c) show simulated realizations of

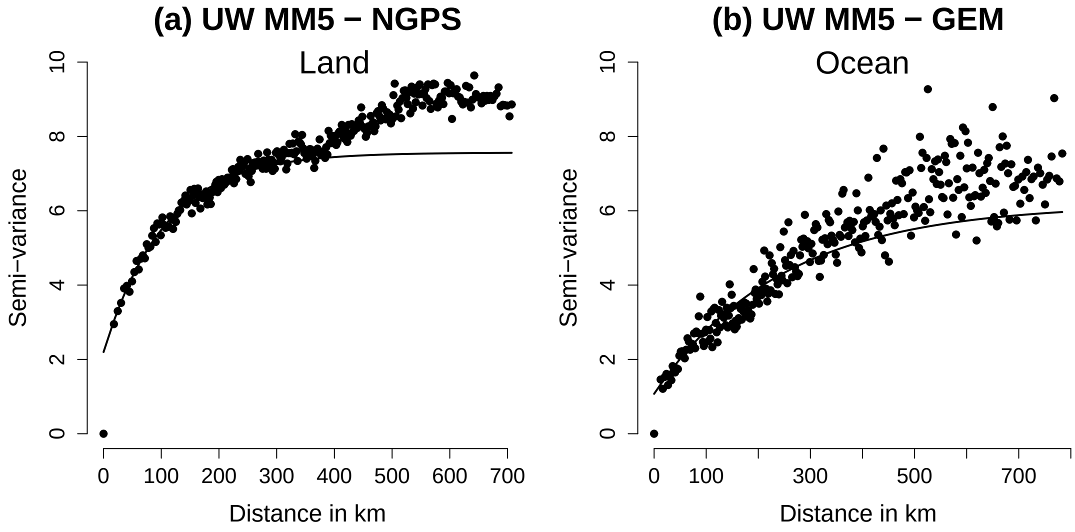

Figure 1: Empirical variograms of 48-h forecast errors for surface temperature over a 25 day training period ending 14 February 2004, using UWME: (a) ngps member on land, (b) cmcg member over ocean.

the continuous and discontinuous components of the forecast error field, respectively. Panel (d) shows the member of the Spatial BMA ensemble as the sum of the three components. Repeating this process, statistical ensembles of any size can be generated.

For each region, we generated a Spatial BMA ensemble of 19 weather fields. These could be displayed in the form of a stamp plot, which we omit. Ensemble forecasts for all types of composite quantities can be derived from the statistical ensemble. For instance, we might be interested in predicting the empirical variogram of the temperature field verifying at 0000 UTC on 16 February 2004. We computed the empirical variogram for each of the 19 members of the Spatial BMA ensemble, using 300 distance bins. At each bin, the minimum and the maximum of the respective 19 values envelop a 95% prediction interval for the verifying variogram value, which we computed from the observed temperature field. Figure 4 shows the results of this experiment. The prediction intervals generally cover the verifying empirical variogram values.

# 4 Verification results

In calendar year 2004, the 0000 UTC cycle for the 12-km domain of the eight-member University of Washington mesoscale ensemble (UWME) was operational on 245 days. For each day, we fitted BMA, GOP and Spatial BMA models for 48-h forecasts of surface (2-m) temperature over the North American Pacific Northwest, separately on land and over the ocean, and using a sliding 25-day training period. We then generated Independent BMA,

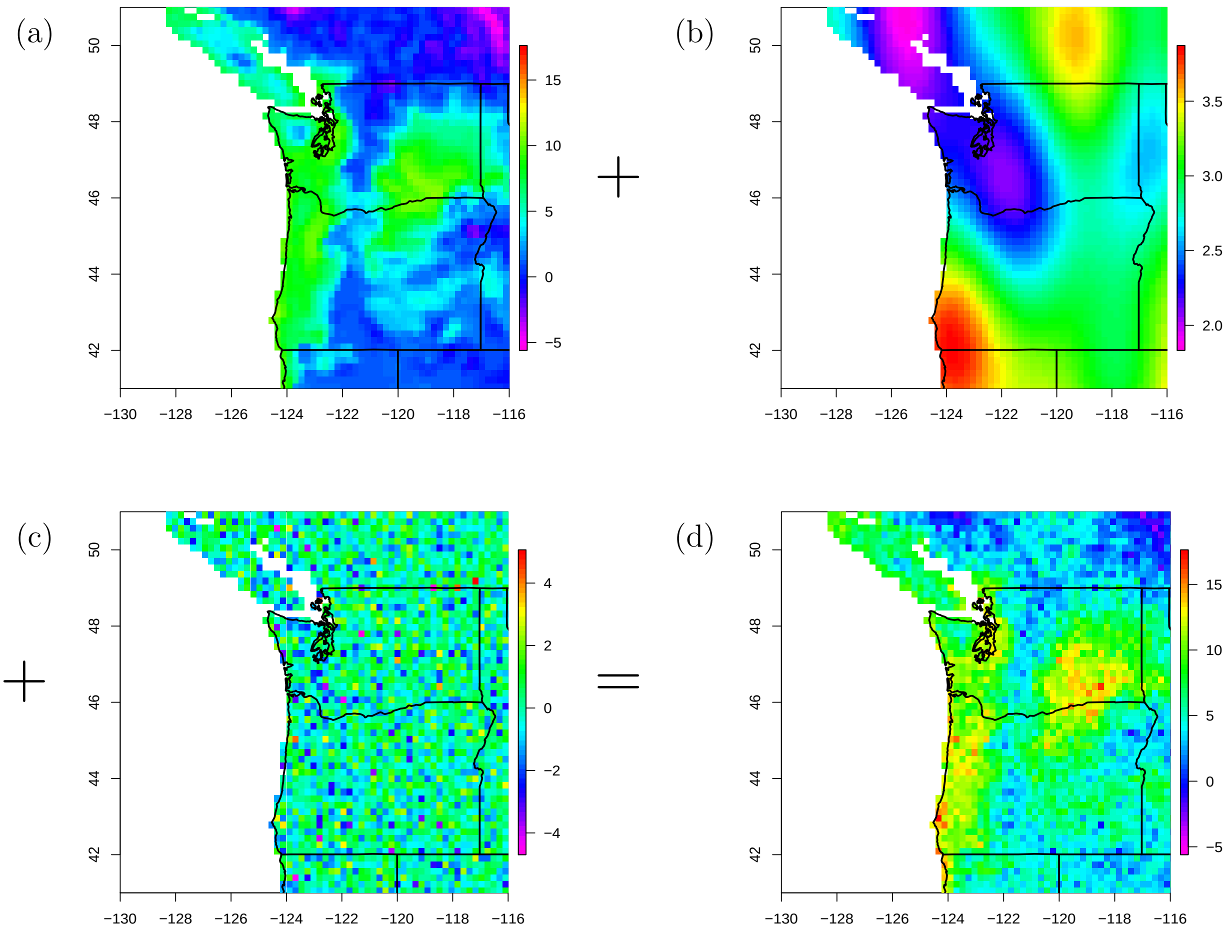

Figure 2: A member of the Spatial BMA ensemble for 48-h forecasts of surface temperature over the land portion of the Pacific Northwest, initialized at 0000 UTC on 14 February 2004: Adding (a) the bias-corrected UWME ngps weather field forecast, (b) the continuous, and (c) the discontinuous component of the simulated forecast error field, we obtain (d) a member of the Spatial BMA ensemble. Note that different color scales are used in the four panels to make it easier to see the patterns in each one.

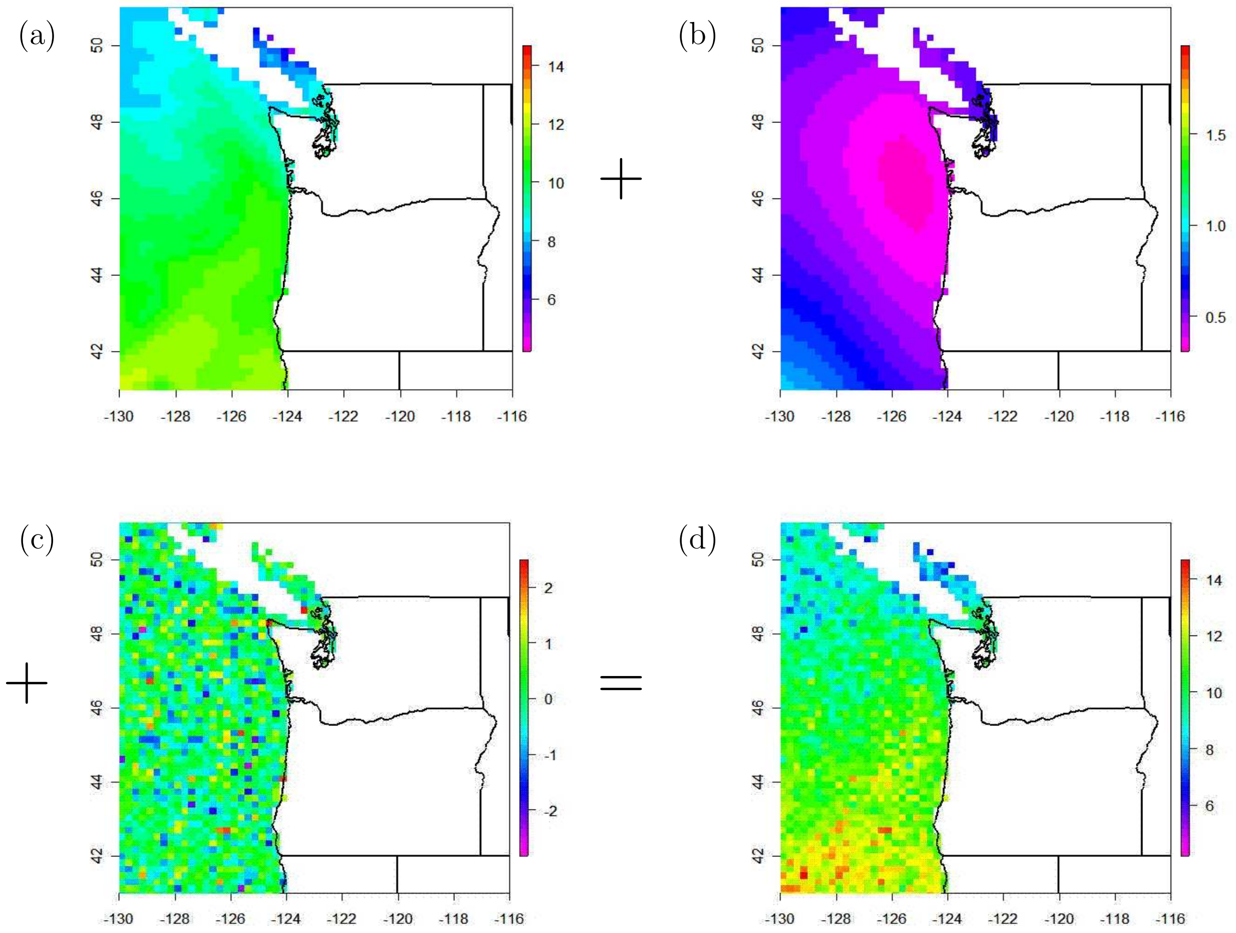

Figure 3: Same as Figure 2, but for the UWME cmcg member and over the Pacific Ocean.

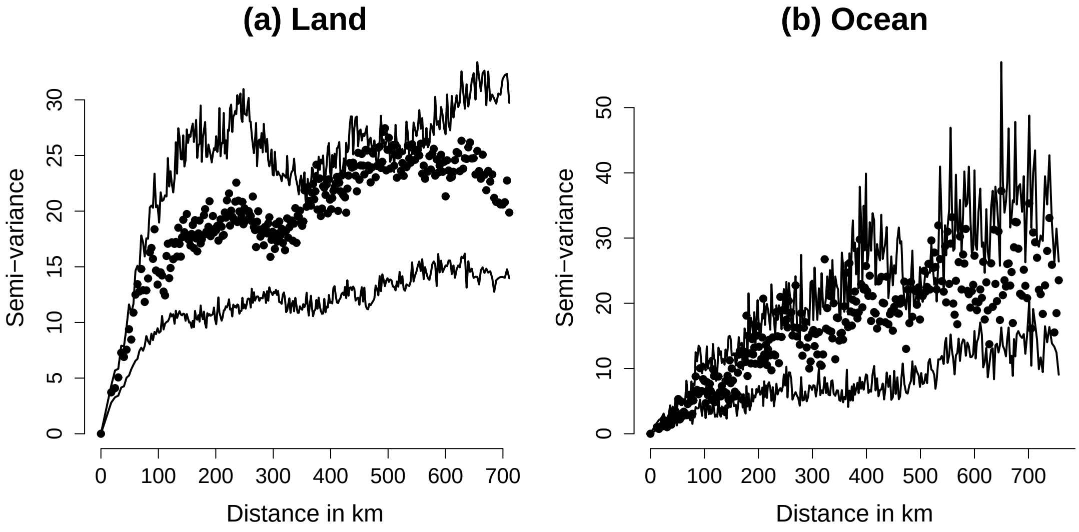

Figure 4: Empirical variogram values (dots) for the verifying surface temperature field at 0000 UTC on 16 February 2004, and pointwise minimum and maximum of the empirical variogram values (lines) from the 19-member Spatial BMA weather field ensemble (a) on land and (b) over ocean.

GOP and Spatial BMA forecast ensembles for each day. The Independent BMA ensembles were created by sampling from the univariate original BMA predictive PDFs at each location separately, mistakenly assuming spatial independence of the forecast error fields. The GOP ensembles were based on the UWME ukmo model, which had the best aggregate performance among the ensemble member models, both on land and over the ocean. In the interest of a fair comparison to the eight-member UWME, our GOP, Independent BMA and Spatial BMA ensembles also had eight members only. However, the statistical ensembles allow for ensembles of any size, and larger ensembles frequently show better verification results.

We now assess and rank the performance of the UWME, GOP, Independent BMA and Spatial BMA ensembles, emphasizing spatial and composite quantities. On average, observations of surface temperature were available at 761 stations on land and 196 stations over the Pacific Ocean. We verified bilinearly interpolated ensemble forecasts against the temperature observations.

In contrast to the statistical ensembles, UWME is not designed to take instrument and representativeness errors into account. Hence, we consider a fifth ensemble, which we denote as UWME + Noise. To create the UWME + Noise ensemble, we added Gaussian noise to each of the eight UWME members, at each site independently, and with mean zero and a variance that equals the estimated nugget effect, $\rho_k^2$, for the respective member model.

Table 3: Mean absolute error (MAE) and average continuous ranked probability score (CRPS) for 48-h forecasts of surface temperature over the Pacific Northwest in 2004, in degrees Celsius.

| | Land | | Ocean | |
|---|---|---|---|---|
| Ensemble | MAE | CRPS | MAE | CRPS |
| UWME | 2.94 | 2.58 | 2.44 | 2.12 |
| UWME + Noise | 2.94 | 2.23 | 2.44 | 1.89 |
| GOP | 2.71 | 2.13 | 2.35 | 1.82 |
| Independent BMA | 2.70 | 1.95 | 2.35 | 1.72 |
| Spatial BMA | 2.70 | 1.95 | 2.35 | 1.72 |

## 4.1 Temperature forecasts at individual sites

We begin by assessing surface temperature forecasts at individual sites. For forecasts at single sites, Spatial BMA and Independent BMA are equivalent; hence, the results for the two ensembles are essentially identical, with any differences due to chance variability in the generation of the ensemble members. All verifications statistics were spatially and temporally composited over the Pacific Northwest and calendar year 2004.

Table 3 shows the mean absolute error (MAE) and the average continuous ranked probability Score (CRPS; Hersbach 2000; Gneiting et al. 2005; Wilks 2006, Section 7.5.1) for the various ensemble methods. The MAE assesses the accuracy of the respective deterministic forecasts. The UWME and UWME + Noise deterministic forecast is the raw ensemble mean; for the GOP method this is the bias-corrected UWME ukmo forecast; and for the Independent BMA and Spatial BMA techniques this is a weighted average of the bias-corrected ensemble member forecasts. The CRPS is a scoring rule for predictive PDFs that addresses calibration as well as sharpness, and is proper, that is, discourages hedging. The CRPS generalizes the absolute error, to which it reduces for deterministic forecasts; it is also reported in degrees Celsius, and average CRPS values can be directly compared to the MAE (Gneiting et al. 2005). A clear rank order can be observed, in that the BMA ensembles showed substantially lower CRPS values than the GOP ensemble, followed by the UWME + Noise and UWME ensembles.

To assess the calibration of the ensemble forecasts, we use the verification rank histogram (Anderson 1996; Talagrand et al. 1997; Hamill and Colucci 1997; Hamill 2001). Figure 5 shows the histograms for the various ensembles. We also computed the respective discrepancy from uniformity,

$$D = \sum_{j=1}^{K+1} \left| p_j - \frac{1}{K+1} \right|, \tag{10}$$

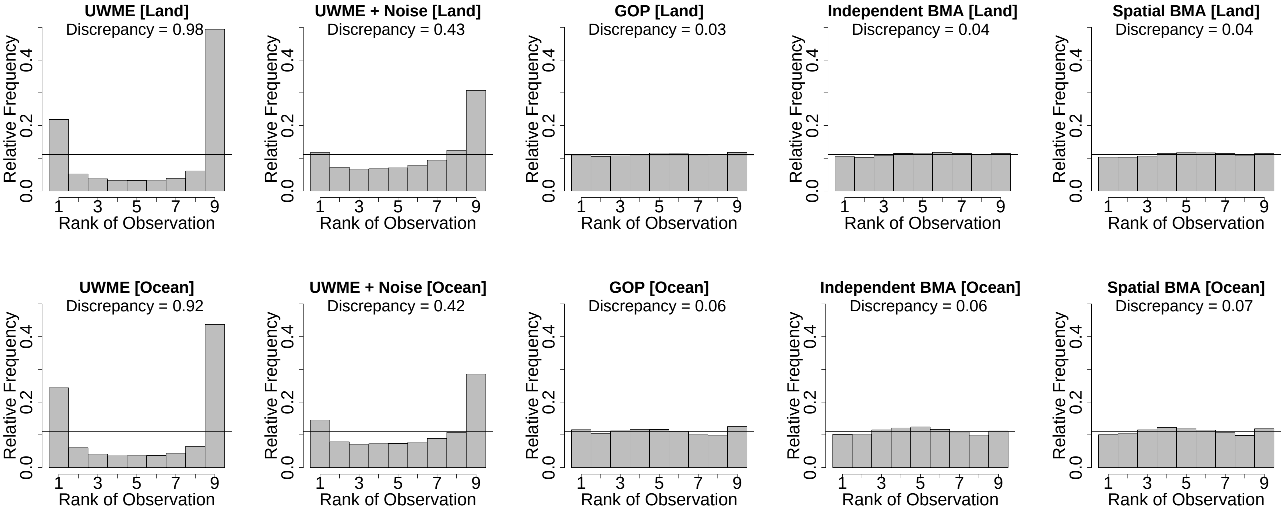

Figure 5: Verification rank histograms for 48-hour forecasts of surface temperature over the Pacific Northwest in 2004 for individual stations on land (top row) and over the ocean (bottom row).

where $K = 8$ is the number of ensemble members and $p_j$ is the observed relative frequency of rank $j$. The smaller the discrepancy, the smaller the deviation from a uniform rank histogram, and the better the calibration. In both domains, the UWME and UWME + Noise ensembles were underdispersive, while the GOP, Independent BMA and Spatial BMA ensembles had rank histograms that were nearly uniform.

## 4.2 Temperature field forecasts

To assess the calibration of the ensembles as weather field forecasts, rather than forecasts of weather quantities at individual sites, we use a variant of the verification rank histogram that is tailored to this task, namely the minimum spanning tree (MST) rank histogram (Smith and Hansen 2004; Wilks 2004). The details of the methodology are described in the aforementioned references. Here, suffice it to say that an MST rank $k \in \{1, \ldots, K+1\}$ is computed based on each day's ensemble of weather field forecasts and the verifying weather field. This yields 245 MST ranks for each of the five ensemble techniques, and the respective histogram is uniform if the ensemble is calibrated. Figure 6 shows the MST rank histograms for the UWME, UWME + Noise, GOP, Independent BMA and Spatial BMA weather field ensembles, separately on land and over the ocean, along with the discrepancy (10) that measures the departure from uniformity. The UWME, UWME + Noise and GOP weather field ensembles were severely underdispersive. The Independent BMA ensemble also was underdispersive, but to a lesser extent. The MST rank histograms for the Spatial BMA weather field ensemble departed the least from uniformity. The difference between the GOP

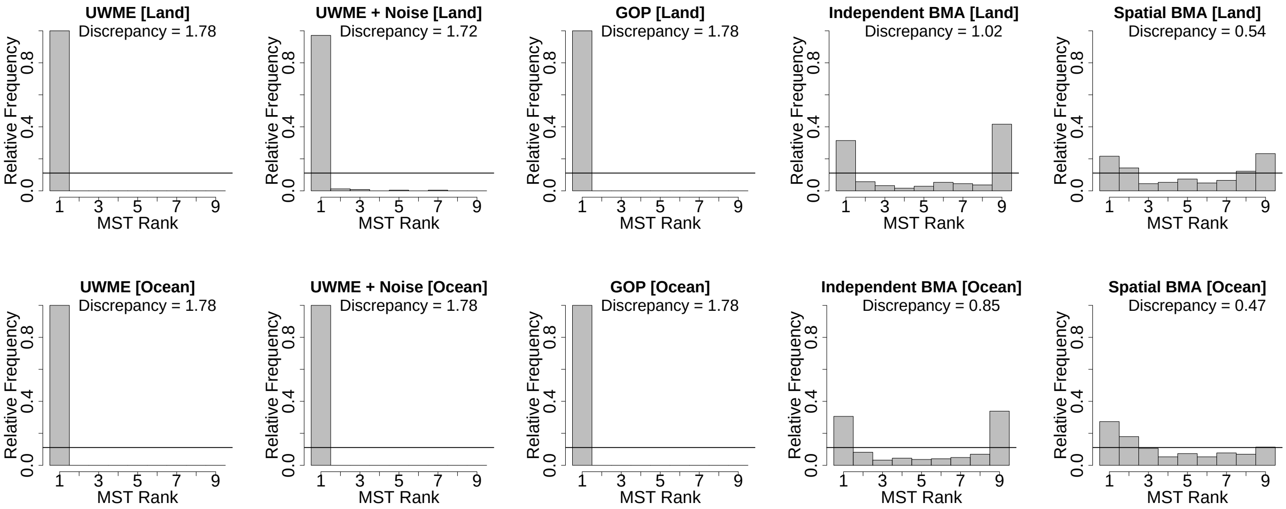

Figure 6: Minimum spanning tree (MST) rank histograms for 48-h weather field forecasts of surface temperature over the Pacific Northwest in 2004 on land (top row) and over the ocean (bottom row).

Table 4: Coverage of nominal 77.8% prediction intervals for variogram values.

| Ensemble | Land | Ocean |
|---|---|---|
| UWME | 20.5 | 28.8 |
| UWME + Noise | 36.3 | 42.8 |
| GOP | 56.6 | 58.7 |
| Independent BMA | 30.9 | 46.3 |
| Spatial BMA | 60.1 | 57.1 |

and Spatial BMA ensembles corroborates the widely held perception that it is advantageous to take account of the flow-dependent information contained in the dynamical ensemble.

As an alternative approach to the spatial verification of ensembles of weather field forecasts, we repeated the variogram computations in Figure 4 for the 245 available days in 2004 and the five types of weather field ensembles. Each eight-member ensemble supplies nominal $\frac{7}{9} \times 100\% = 77.8\%$ prediction intervals for variogram values computed from the verifying temperature field. Table 4 shows the empirical coverage of the prediction intervals when composited over the 245 days and 300 distance bins. For all five types of ensembles, the empirical coverage was lower than desired, but the coverage for the GOP and Spatial BMA ensembles was closest to nominal.

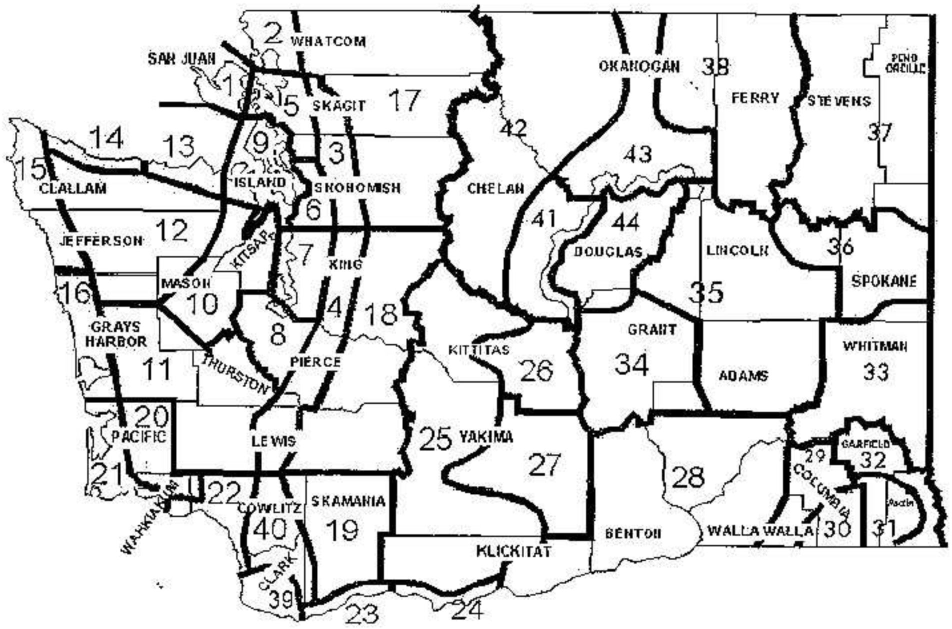

Figure 7: National Weather Service forecast zones in the state of Washington, bordered by the Pacific Ocean to the west, and British Columbia, Idaho and Oregon to the north, east and south (www.atmos.washington.edu/data/images/zone.gif).

## 4.3 Average temperature in National Weather Service forecast zones

Spatial correlations play crucial roles in the prediction of a number of composite quantities. Here, we present verification results for ensemble forecasts of spatial averages of temperature. Figure 7 shows the 44 National Weather Service (NWS) forecast zones in the state of Washington. For each zone and each day, we considered ensemble forecasts of average surface temperature, understood as the mean of the temperature observations at the stations within the zone.

Figure 8 summarizes verification statistics for the various types of eight-member ensembles in the 44 zones. The performance of the GOP ensemble was almost identical to that of the Spatial BMA ensemble, and we omit the respective results. Panel (a) shows the discrepancy (10) that measures the departure of the verification rank histogram from uniformity. In almost all zones, the Spatial BMA ensemble showed the lowest discrepancy. Figure 9 illustrates this for forecast zone 7, which has one of the highest numbers of stations and contains the city of Seattle. Both UWME, UWME + Noise and the Independent BMA ensemble were underdispersive, while the GOP and Spatial BMA ensembles had verification rank histograms that were similar to each other and close to being uniform. The underdispersion of the Independent BMA ensemble is not surprising, in that the assumption of spatial

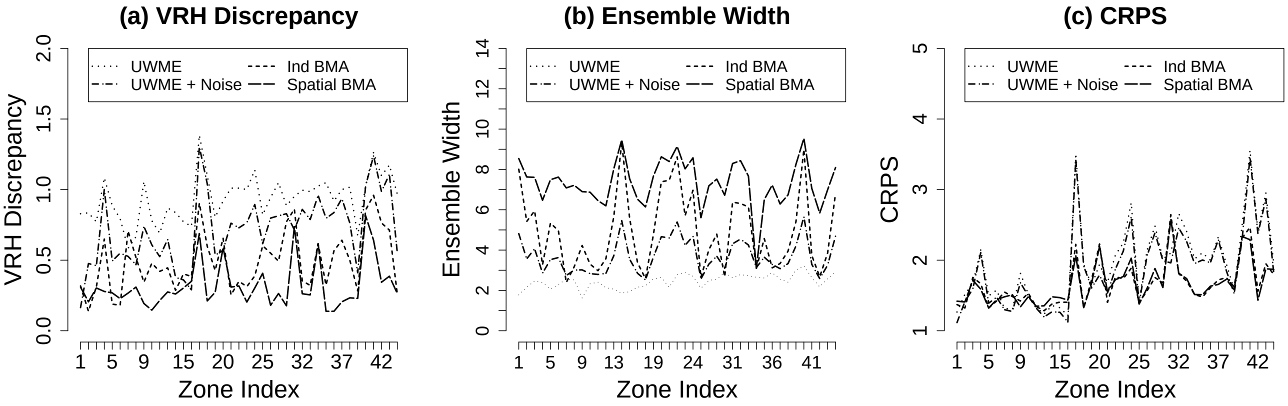

Figure 8: Verification statistics for 48-h forecasts of average surface temperature in NWS forecast zones in 2004. (a) Verification rank histogram discrepancy. (b) Mean ensemble width in degrees Celsius. (c) Mean CRPS value in degrees Celsius.

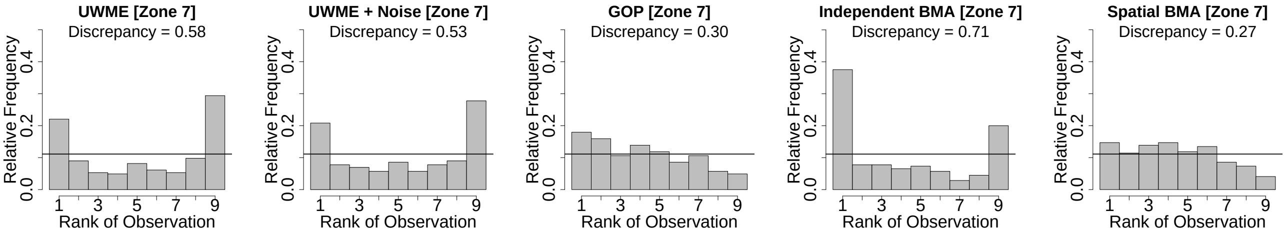

Figure 9: Verification rank histograms for 48-h forecasts of average surface temperature in NWS forecast zone 7 in 2004.

independence of forecast errors implies an underestimation of the variance of temperature averages.

Panel (b) in Figure 8 shows the average range of the forecast ensemble for the various types of ensembles. The range quantifies the sharpness of the predictive distributions and is simply the difference between the maximum and the minimum of the eight ensemble values. The UWME had the sharpest predictive distributions, but it was underdispersive, and therefore uncalibrated. A similar comment applies to the Independent BMA ensemble. The Spatial BMA ensemble was the least sharp, but it was better calibrated than the other types of ensembles. Finally, to assess calibration and sharpness simultaneously, panel (c) shows the respective aggregate continuous ranked probability score (CRPS) values. Despite being sharpest, the UWME generally had the highest, least desirable CRPS values. The Independent BMA and the Spatial BMA ensembles had CRPS values that were lower, and quite similar to each other, even though the ensembles behaved quite differently in terms of calibration and sharpness. Gneiting et al. (2003) argued that the goal of probabilistic forecasting is to maximize sharpness under the constraint of calibration, and from this perspective, the

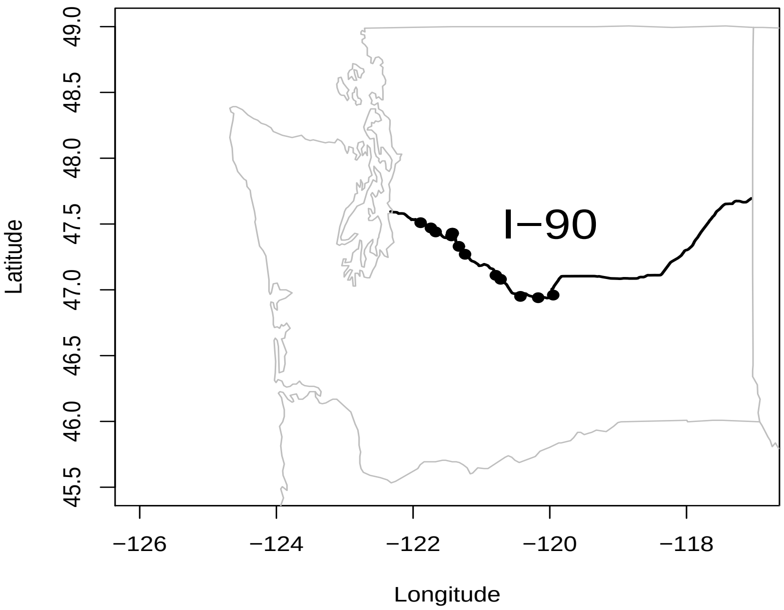

Figure 10: Meteorological stations along the Cascades corridor of Interstate 90.

performance of the Spatial BMA ensemble is superior.

## 4.4 Minimum temperature along Interstate 90

We now present verification results for another composite quantity: minimum temperature along the Interstate 90 Mountains to Sound Greenway, Washington's primary east-west bound highway. Accommodating 20 million travelers annually, Interstate 90 crosses the Cascade Mountains in a dramatic mountain landscape with substantial altitude differentials. Accurate and reliable forecasts of minimum temperature are critical to highway maintenance operations.

Figure 10 shows the locations of 13 meteorological stations along the Cascades section of Interstate 90, some of which are very near each other. We consider ensemble forecasts of the minimum temperature among these 13 stations. The UWME forecasts were available on the 12-km model grid and were bilinearly interpolated to the observation locations. However, Interstate 90 and the meteorological stations are generally located at lower altitudes, while the surrounding gridpoints are at higher altitudes. Hence, altitude is a critical consideration, and we applied a standard lapse rate correction of 0.65 degrees Celsius per 100 m to all five types of forecast ensembles.

Figure 11 shows verification rank histograms for the eight-member UWME, UWME + Noise, GOP, Independent BMA and Spatial BMA forecast ensembles. The UWME and UWME + Noise ensembles were underdispersive. The Independent BMA ensemble was

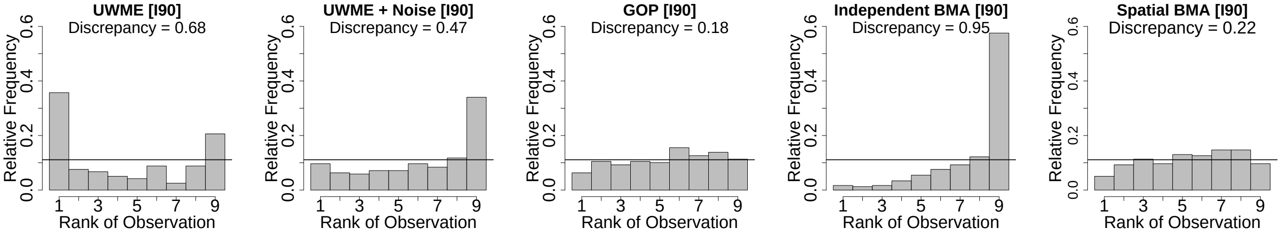

Figure 11: Verification rank histograms for 48-h forecasts of minimum temperature along Interstate 90.

Table 5: Verification rank histogram discrepancy, mean ensemble width and mean CRPS value for ensemble forecasts of minimum temperature along Interstate 90. The unit for the ensemble range and CRPS is degrees Celsius.

| Ensemble | Discrepancy | Range | CRPS |
|---|---|---|---|
| UWME | 0.68 | 2.76 | 1.55 |
| UWME + Noise | 0.47 | 4.37 | 1.67 |
| GOP | 0.18 | 7.75 | 1.53 |
| Independent BMA | 0.95 | 6.48 | 2.76 |
| Spatial BMA | 0.22 | 8.04 | 1.54 |

strongly biased, tending to underestimate the minimum temperature along Interstate 90. Indeed, the minimum of a collection of independent forecasts tends to be smaller than the minimum of a collection of forecasts that are spatially correlated. The GOP and Spatial BMA ensembles had rank histograms that were close to being uniform. Table 5 shows the verification rank histogram discrepancy, the mean ensemble range and the mean CRPS value for the forecast ensembles. The UWME, GOP and Spatial BMA ensembles showed similar CRPS values, thereby illustrating a trade-off between calibration and sharpness. In view of our goal of maximizing sharpness under the constraint of calibration, we contend that the GOP and Spatial BMA ensembles are preferable for most users.

# 5 Discussion

We have introduced the Spatial BMA method, a statistical postprocessing technique for calibrating forecast ensembles of whole weather fields simultaneously. Spatial BMA generalizes and combines Bayesian model averaging (BMA) and the geostatistical output perturbation (GOP) technique, and it honors ensemble as well as spatial statistical information. The Spatial BMA predictive PDF for the weather field is a weighted average of multivariate normal PDFs centered at bias-corrected members of the dynamical forecast ensemble. At any single location, Spatial BMA reduces to the original BMA technique. It is computationally inexpensive and can be used to generate statistical ensembles of any size.

In experiments with the University of Washington mesoscale ensemble, the Spatial BMA ensemble compared favorably to the raw dynamical ensemble, the raw ensemble with added observational noise, the GOP ensemble and the Independent BMA ensemble. In particular, the minimum spanning tree histogram, a key tool in assessing the calibration of ensembles of weather field forecasts (Smith and Hansen 2004; Wilks 2004), was closest to being uniform for the Spatial BMA ensemble. For forecasts of composite quantities, such as temperature averages over NWS forecast zones and minimum temperature along the Cascades corridor of Interstate 90, the GOP ensemble and the Spatial BMA ensemble showed similar performances, and outperformed the other types of ensembles. While our experiments were with surface temperature fields, Spatial BMA in its present form applies to all weather variables with forecast error distributions that are approximately Gaussian, including sea-level pressure. Further research is needed to extend Spatial BMA to other weather variables, such as precipitation or wind speed. Sloughter et al. (2006) presented a non-Gaussian version of BMA that yields calibrated quantitative probabilistic precipitation forecasts at individual sites, but not for weather fields.

There are several directions into which the Spatial BMA technique could be developed. One of them is bias correction. In the current implementation, we use a simple linear bias correction that does not take altitude, land use, latitude, longitude, or distance from the ocean into account. More sophisticated regression based bias removal techniques might include some or all of these quantities as predictor variables. Another possibility is to use a nearest neighbor approach based on distance, altitude and land use categories. This approach is currently under investigation in the group of Clifford F. Mass in the Department of Atmospheric Sciences at the University of Washington, and initial results are encouraging.

In modeling the covariance structure of the forecast error fields, we used a stationary and isotropic, exponential correlation function. There are several ways in which more complex, and potentially more realistic, covariance structures could be employed. Stationary and isotropic correlation functions that are more versatile than an exponential function are available (Gneiting 1999). Anisotropic covariance structures could also be employed; however, in the case of surface temperature over the Pacific Northwest, Gel et al. (2004b) did not find any significant differences between longitudinal and latitudinal empirical variograms of the forecast error fields. Finally, nonstationary covariance models could be used. In our experiments, we dealt with nonstationarities between the land and the ocean by fitting and generating two distinct Spatial BMA ensembles, each of which used a stationary and isotropic covariance structure. This was a fairly simple way to resolve nonstationarities, and yet produced good results. The methods of Paciorek and Schervish (2005) could be used to fit valid covariance structures that are stationary on homogeneous domains, yet nonstationary globally, thereby allowing for the generation of a single Spatial BMA ensemble over all domains simultaneously, without incurring discontinuities along the boundaries.

An issue not explicitly considered in the Spatial BMA approach is that of phase or displacement errors. These could perhaps be addressed by partitioning the errors of the ensemble member weather field forecasts into displacement, distortion, amplitude and residual fields, as in Du et al. (2000), and applying the Spatial BMA technique to the residual component only, while developing parametric statistical models for displacement, distortion and amplitude errors. This would be an interesting avenue for future research, with potential rewards in the form of sharper yet calibrated forecast PDFs, but may require impractically large sets of training data.

A characteristic feature of the Spatial BMA and GOP ensemble member fields is an increase in roughness, when compared to weather fields generated by numerical weather prediction models. This stems from Spatial BMA aiming to reproduce the spatial structure of weather observations, including instrument and representativeness errors. The Spatial BMA forecast error fields decompose into a smooth, continuous component and a discontinuous component, with the latter taking the instrument and representativeness errors into account. The discontinuous component can be ignored, if desirable, and the Spatial BMA technique can be implemented by adding the bias-corrected weather field forecast (panel (a) in Figures 2 and 3) and the continuous component of the simulated forecast error field (panel (b)) only. This is an implementation decision that needs to be made depending on the prediction problem at hand.

Another issue that calls for discussion is the choice of the training period. In the current implementation, we use forecast and observation data from a sliding window consisting of the 25 most recent days available to estimate the Spatial BMA parameters. This allows the method to adapt rapidly to seasonal changes in the atmosphere as well as changes in the design of the ensemble, but limits the availability of training data. A potential alternative is to also use training data from the same season in previous years, and this could be done using ensemble reforecasts, as proposed by Hamill et al. (2004). However, reforecasts put high demands on computational and human resources, and they were not available to us.

We close by comparing Spatial BMA to other ensemble postprocessing techniques. Wilks (2002) proposed to fit mixtures of multivariate normal densities to ensemble forecasts of multivariate weather quantities. This resembles the Spatial BMA technique, but does not take bias and calibration adjustments into account. Roulston and Smith (2003) led the way in proposing to combine statistical and dynamical ensembles, and suggested the use of hybrid ensembles, in which the members of the dynamical ensemble are dressed with errors drawn from an archive of best member errors. A difficulty in this approach is the identification of best members. Wang and Bishop (2005) showed that under a wide range of scenarios the best member dressing method fails to be calibrated. They proposed a modified dressing technique, in which statistical perturbations are generated, with flexible covariance structures that are estimated from training data. This is similar to the Spatial BMA technique, in that the Wang and Bishop (2005) predictive PDF is also a weighted average of multivariate normal densities, each centered at a bias-corrected member of the dynamical forecast ensemble, but the weights are all equal and do not depend on the member's skill. Fortin and Favre (2006) proposed dressing kernels for exchangeable ensembles that depend on the rank of the member within the ensemble. This allows for differing weights, but it is often difficult to reliably estimate the weights and the dressing kernels.

# Acknowledgements

We thank Jeff Baars, Fadoua Balabdaoui, F. Anthony Eckel, Yulia Gel, Eric P. Grimit, Nicholas A. Johnson, Clifford F. Mass, J. McLean Sloughter and Patrick Tewson, for sharing code, comments and/or data. This research was supported by the DoD Multidisciplinary University Research Initiative (MURI) program administered by the Office of Naval Research under Grant N00014-01-10745.

# References

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate*, **9**, 1518–1530.

Buizza, R., 1997: Potential forecast skill of ensemble prediction and spread and skill distribution of the ECMWF ensemble prediction system. *Mon. Wea. Rev.*, **125**, 99–119.

Byrd, R. H., P. Lu, J. Nocedal and C. Zhu, 1995: A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.*, **16**, 1190–1208.

Chilès, J.-P. and P. Delfiner, 1999: *Geostatistics: Modeling Spatial Uncertainty.* Wiley, 695 pp.

Cressie, N. A. C., 1993: *Statistics for Spatial Data.* Wiley, revised edition, 900 pp.

Dempster, A. P., N. M. Laird and D. B. Rubin, 1977: Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B*, **39**, 1–39.

Du, J., S. Mullen and F. Sanders, 2000: Removal of distortion error from an ensemble forecast. *Mon. Wea. Rev.*, **128**, 3347–3351.

Eckel, F. A. and C. F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting*, **20**, 328–350.

Epstein, E. S., 1969: Stochastic dynamic prediction. *Tellus*, **21**, 739–759.

Fortin, V. and A.-C. Favre, 2006: Taking into account the rank of a member within the ensemble for probabilistic forecasting based on the best member method. in *Preprints, 18th Conference on Probability and Statistics in the Atmospheric Sciences*, Atlanta, GA.

Gel, Y., A. E. Raftery and T. Gneiting, 2004a: Calibrated probabilistic mesoscale weather field forecasting: The geostatistical output perturbation (GOP) method (with discussion). *J. Amer. Stat. Assoc.*, **99**, 575–588.

Gel, Y., A. E. Raftery, T. Gneiting and V. J. Berrocal, 2004b: Calibrated probabilistic mesoscale weather field forecasting: The geostatistical output perturbation (GOP) method: Rejoinder. *J. Amer. Stat. Assoc.*, **99**, 589–590.

Glahn, H. R. and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211.

Gneiting, T., 1999: Correlation functions for atmospheric data analysis. *Quart. J. Roy. Meteor. Soc.*, **125**, 2449–2464.

Gneiting, T. and A. E. Raftery, 2005: Weather forecasting using ensemble methods. *Science*, **310**, 248–249.

Gneiting, T., A. E. Raftery, F. Balabdaoui and A. H. Westveld, 2003: Verifying probabilistic forecasts: Calibration and sharpness. in *Proc. Workshop on Ensemble Forecasting*, Val-Morin, Quebec, Canada. [Available at www.cdc.noaa.gov/~hamill/ef_workshop_2003_schedule.html].

Gneiting, T., A. E. Raftery, A. H. Westveld and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118.

Gneiting, T., H. Ševčíková, D. B. Percival, M. Schlather and Y. Jiang, 2006: Fast and exact simulation of large Gaussian lattice systems in $\mathbb{R}^2$: Exploring the limits. *J. Comput. Graph. Stat.*, in press.

Grimit, E. P. and C. F. Mass, 2002: Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest. *Wea. Forecasting*, **17**, 192–205.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560.

Hamill, T. M. and S. J. Colucci, 1997: Verification of Eta-RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125**, 1312–1327.

Hamill, T. M., J. S. Whitaker and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434–1447.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570.

Hoeting, J. A., D. M. Madigan, A. E. Raftery and C. T. Volinsky, 1999: Bayesian model averaging: A tutorial (with discussion). *Stat. Sci.*, **14**, 382–401, A corrected version with typos corrected is available at www.stat.washington.edu/www/research/online/hoeting1999.pdf.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.*, **124**, 1225–1242.

Houtekamer, P. L. and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.*, **126**, 796–811.

Houtekamer, P. L. and H. L. Mitchell, 2001: A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, **129**, 123–137.

Ihaka, R. and R. Gentleman, 1996: R: A language for data analysis and graphics. *J. Comput. Graph. Stat.*, **5**, 299–314.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418.

Molteni, F., R. Buizza, T. N. Palmer and T. Petroliagis, 1996: The ECMWF ensemble system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119.

Paciorek, C. J. and M. J. Schervish, 2005: Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, under review.

Palmer, T. N., 2002: The economic value of ensemble forecasts as a tool for assessment: From days to decades. *Quart. J. Roy. Meteor. Soc.*, **128**, 747–774.

Raftery, A. E., T. Gneiting, F. Balabdaoui and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126**, 649–667.

Roulston, M. S. and L. A. Smith, 2003: Combining dynamical and statistical ensembles. *Tellus*, **55A**, 16–30.

Scherrer, S. C., C. Appenzeller, P. Eckert and D. Cattani, 2004: Analysis of the spread-skill relations using the ECMWF ensemble prediction system over Europe. *Wea. Forecasting*, **19**, 552–565.

Schlather, M., 2001: Simulation and analysis of random fields. *R News*, **1(2)**, 18–20.

Sloughter, J. M., A. E. Raftery and T. Gneiting, 2006: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.*, submitted.

Smith, L. A. and J. A. Hansen, 2004: Extending the limits of ensemble forecast verification with the minimum spanning tree. *Mon. Wea. Rev.*, **132**, 1522–1528.

Stensrud, D. J., H. E. Brooks, J. Du, M. S. Tracton and E. Rogers, 1999: Using ensembles for short-range forecasting. *Mon. Wea. Rev.*, **127**, 433–446.

Talagrand, O., R. Vautard and B. Strauss, 1997: Evaluation of probabilistic prediction systems. in *Proc. ECMWF Workshop on Predictability*, pp. 1–25, Reading, UK.

Toth, Z. and E. Kalnay, 1993: Ensemble forecasting at the NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330.

Wandishin, M. S., S. L. Mullen, D. J. Stensrud and H. E. Brooks, 2001: Evaluation of a short-range multimodel ensemble system. *Mon. Wea. Rev.*, **129**, 729–747.

Wang, X. and C. H. Bishop, 2005: Improvement of ensemble reliability with a new dressing kernel. *Quart. J. Roy. Meteor. Soc.*, **131**, 965–986.

Wilks, D. S., 2002: Smoothing forecast ensembles with fitted probability distributions. *Quart. J. Roy. Meteor. Soc.*, **128**, 2821–2836.

Wilks, D. S., 2004: The minimum spanning tree histogram as a verification tool for multidimensional ensemble forecasts. *Mon. Wea. Rev.*, **132**, 1329–1340.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. Elsevier Academic Press, second edition, 627 pp.

Wood, A. T. A. and G. Chan, 1994: Simulation of stationary Gaussian processes in $[0,1]^d$. *J. Comput. Graph. Stat.*, **3**, 409–432.